Cyberspace Operations Functional Capability Reference Architecture from Document Text

Allen Moulton
Research Scientist
SSRC
Massachusetts Institute of Technology

Stuart. E. Madnick
Professor
Sloan School of Management
Massachusetts Institute of Technology

Nazli Choucri
Professor
Political Science Department
Massachusetts Institute of Technology

September 2020

Abstract

The COMET project applies structured text analysis, semantic similarity and ontology learning theory, along with NLP to investigate automated and semi-automated methods for extracting knowledge from text policy documents and transforming that knowledge into a structured form for use in a Functional Capability Reference Architecture (FCRA) for cyberspace operations. Progress and results are reported.

Keywords

Cybersecurity, Reference Architecture, NLP, Semantics, Defense.

Cyberspace Operations Functional Capability Reference Architecture from Document Text

Allen Moulton, Stuart Madnick, Nazli Choucri

Working Paper CISL# 2020-24

September 2020

Cybersecurity Interdisciplinary Systems Laboratory (CISL) Sloan School of
Management, Room E62-422 Massachusetts Institute of Technology
Cambridge, MA 02142

Cyberspace Operations Functional Capability Reference Architecture from Document Text

Allen Moulton, Stuart Madnick and Nazli Choucri

Keywords: Cybersecurity, Reference Architecture, NLP, Semantics, Defense

The COMET project applies structured text analysis, semantic similarity and ontology learning theory, along with NLP to investigate automated and semi-automated methods for extracting knowledge from text policy documents and transforming that knowledge into a structured form for use in a Functional Capability Reference Architecture (FCRA) for cyberspace operations. Progress and results are reported.

Cyber-physical systems are increasingly significant to modern life. In the military, the advent of net-centric systems means that virtually all operations critically depend on computers and networks (Williams, 2014). The operation of the electric power grid is moving in the same direction (GAO, 2019) as are most other industries. As Choucri and Clark (2019) document, cyberspace has also become increasingly intertwined in international politics. To make cyber-physical systems more effective and to protect from threats that put critical services at risk, organizations rely on policy documents which are written from different perspectives often using different terminology. In many cases, terminology is metaphorical such as maneuver, attack and defense, which draw on analogies to concepts from physical domain military operations. A FCRA will support knowledge transfer across different subject areas and organizations by harmonizing and clarifying concepts (Cloutier et al., 2010).

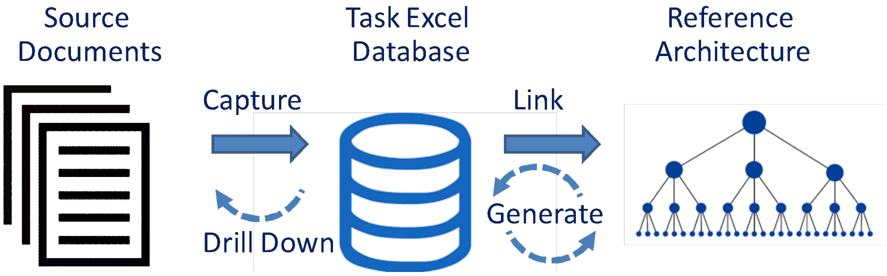

Figure 1. Stages of information flow in COMET Project

COMET was structured into the three decoupled stages shown in Figure 1. COMET collected a sample of 5,915 pages in sources ranging from unclassified, public release military doctrine (http://www.jcs.mil/Doctrine/), NIST cybersecurity policy (e.g., NIST, 2018) and other similar separately-authored sources. We included the full spectrum of Cyberspace Operations from actions to secure friendly systems by removing vulnerabilities (security), to active measures to respond to intrusions or attacks (defense), to intelligence gathering about adversary networks and systems (exploitation), through taking actions against adversary systems (attack). Even though most actors would not be allowed to take actions on the aggressive end of the spectrum, the ubiquity and mirror image nature of cyberspace makes it important to understand the full spectrum in order prepare to defend against offensive actions when they are used by an adversary. For example, the MITRE ATT&CK frameworks model targeting and attack actions to help cybersecurity managers better understand threats by interpreting observed anomalies from the point of view of an adversary (http://attack.mitre.org).

COMET focuses on collecting and analyzing information about Tasks, which are the central element of the functional capability concept (Wißotzki, 2018). The Task describes a function -- *what* action is to be done, but not *how*, *why*, *where*, *when*, or *by whom*. Because the Task is central, a capability architecture will have the same structure as its underlying task architecture. Task definitions must be filtered out of the bulk of other information in the policy documents, such as the how, why, and by whom aspects of capabilities, as well as general background information, organizational recommendations, and desired states for the overall system.

The COMET information flow for building and maintaining the cyberspace operations FCRA begins with crawling documents from the corpus, identifying explicit or implicit mentions of Tasks, which are extracted and captured in textual form in a Task Database. From the sample corpus, we extracted 1499 distinct task definitions. In addition to textual data describing each Task, we collected provenance metadata to allow drilling down back to the segment of the original document where each task was found. We also collected metadata about how the Task was identified and the form of the task text in the document to support design of future automated Task extraction procedures.

In the third stage, we experimented with methods to analyze captured Task data and develop a hierarchical functional taxonomical structure for use in the FCRA. Two constraints guided the design: 1) the relatively small size of the Task database (1499) and the potential size if extended to more documents from a universe that extends to thousands, not millions; and 2) the need to structure the result using categories that are known to end users (e.g., warfighting functions known to military commanders (Williams, 2014).

The taxonomy structure is built both top-down from *a priori* fundamental categories known to users in the domain and bottom-up from clustering Task statements based on semantic similarity (Harispe et al., 2016) and ontology learning (Asim, 2018). We used a SME to build a baseline taxonomy for comparison to automated results. In parallel, we initiated efforts to apply off the shelf NLP and machine learning techniques to structure Tasks into categories. By iterating both SME and automated methods we were able to improve the category structure and refine the ambiguous or abbreviated Task definitions by paraphrasing text from the context where the Task was found.

Although the research is not complete, we have been able to show that extracting Tasks from documents and processing them using formal methods can develop a FCRA. We are currently looking to continue the experiments using more advanced NLP and AI techniques. For example, we want to leverage pretrained word vectors to improve measures of semantic similarity and distance used in taxonomy generation. We also want to work on consolidating semantically overlapping Tasks and extending the taxonomy to an ontology incorporating relations representing other aspects of capabilities. Logic-based AI could also provide value for this type of "data-starved" analysis.

References

Asim, M., Wasim, M., Khan, M. U. G., Mahmood, W, Abbasi, H. (2018). A survey of ontology learning techniques and applications, *Database*, Volume 2018, bay101, https://doi.org/10.1093/database/bay101

Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E. and Bone, M. (2010), The Concept of Reference Architectures. Syst. Engin., 13: 14-27. doi:10.1002/sys.20129.

Choucri, N. and Clark, D. (2019). *International Relations in the Cyber Age: The Co-Evolution Dilemma.* MIT Press (2019).

GAO (2019). *Critical Infrastructure Protection: Actions Needed to Address Significant Cybersecurity Risks Facing the Electric Grid*. Report GAO-19-322 (August 2019).

Harispe, S., Ranwez, S., Janaqi, S. and Montmain, J. (2015). Semantic Similarity from Natural Language and Ontology Analysis. *Synthesis Lectures on Human Language Technologies* 2015 8:1, 1-254.

NIST (2018). *Framework for Improving Critical Infrastructure Cybersecurity*, Version 1.1. (Gaithersburg, MD: April 2018).

Wißotzki M. (2018) The Notion of Capability in Literature. In: Sandkuhl K., Stirna J. (eds) *Capability Management in Digital Enterprises*. Springer. 27-39. https://doi.org/10.1007/978-3-319-90424-5_2.

Williams, B. T. (2014). The Joint Force Commander's Guide to Cyberspace Operations. *Joint Force Quarterly* 73. National Defense University.